한글 문자의 서체 분류

김 삼 수, 김 수 형[1)]

요 약

본 논문에서는 한글 문자의 세리프(serif) 계열과 산세리프(sans-serif) 계열의 분류를 위한 특징을 제안한다. 한글의 서체는 세로획의 시작 부분에 장식 세리프(돌기)가 있는 세리프 계열과 그렇지 않은 산세리프 계열로 나눌 수 있다. 제안하는 한글 문자의 서체 분류 방법은 세리프 형태에서 추출한 특징을 이용하여 세리프 또는 산세리프 클래스로 분류하고, 각 클래스별로 적합한 특징 및 분류기를 학습하여 보다 다양한 서체를 인식하도록 계층적으로 설계한다. 제안한 특징의 유용성을 입증하기 위한 실험은 명조, 바탕, 궁서, 고딕, 돋움, 굴림 서체의 3,000개 낱자 영상에 적용하였다.

1. 서론

우리는 흔히 접할 수 있는 신문, 잡지, 책, 계산서, 전표에서 보고서, 저널, 공문서에 이르기까지 수많은 인쇄문서(printed document)들로부터 정보를 얻고 있다. 따라서 대규모 인쇄문서를 저장, 가공, 검색, 재생산하는 등의 응용분야에서 자동 문서인식 및 정보검색의 요구는 당연한 요구라 할 수 있겠다[1]. 문서인식 및 정보검색의 접근 방법에는 문자인식(OCR) 기술을 이용하거나, 문서의 다양한 메타정보(meta-information)를 추출하여 검색에 적용하는 등의 방법들이 있으며[3], 그리고 많은 다른 연구들 중에서 성능 향상의 해결책의 하나로서 문자나 문서의 다양한 속성 정보를 이용하는 광학 폰트 인식(OFR: Optical Font Recognition)에 대한 연구가 활발히 진행되고 있다[3-16].

이러한 폰트 인식의 다양한 응용분야에도 불구하고 아직까지 한글에 대한 폰트 인식은 활발히 이루지지 못하고 있다. 서체 인식을 위해 문서 영상으로부터 추출한 특징들은 개별적인 문자를 대상으로 하는 지역적 특징(local feature)을 사용하거나, 단어[7], 텍스트 라인[5], 텍스트 블록[12, 14]을 대상으로 하는 전역적 특징(global feature)을 사용한다. 지역적 특징을 사용하는 방법은 비교적 정확한 서체 인식을 수행할 수 있으며, 전역적 특징을 사용하는 방법은 다량의 문서를 신속하게 처리할 수 있는 장점이 있으나, 텍스트의 길이나 영상의 품질에 영향을 많이 받는다[5]. 본 논문에서는 서체에 대한 정확한 정보 추출이 목적이므로 개별적인 문자를 대상으로 특징을 추출한다. 또한, 본 논문에서는 통용되는 문서들에 대한 통계적 조사를 통해 다양한 속성 인식의 필요성을 확인하였고, 제안하는 한글 서체인식 시스템이 어떤 서체를 분류 대상으로 삼아야 하는지에 대해 고찰하였다.

본 논문의 2장에서는 한글 문서에서 사용되는 서체에 대한 자료 조사에 대해 기술하고, 3장에서는 제안하는 한글 문자의 서체 분류 시스템을 설계하며, 4장에서는 한글의 서체 분류를 위한 특징을, 5장에서는 낱자 및 단어 영상에 적용한 실험 결과를 설명한다.

2. 서체 사용에 대한 자료 조사

1) 전남대학교 전산학과 교수

서체 인식의 결과를 문서인식 및 정보검색 등에 응용할 경우, 분류 대상을 어떤 서체들로 할 것인가를 결정하기 위해 실제 흔히 접할 수 있는 600여종의 문서들을 대상으로 자료조사를 수행하였다. 자료조사 대상 문서들은 다음과 같이 요약할 수 있다.

- 각종 보고서 및 저널 등의 논문 100여종 (신명조체 사용, 크기 9 - 16)
- 회사 등에서 발행하는 영수증 및 기안 문서 120여종 (굴림, 돋움, 명조체 사용)
- 관공서 등에서 작성되는 공문서 200여종 (신명조, 고딕, 굴림, 고딕체 사용)
- 각종 세무 및 보험고지서 영수증, 전표, 우편봉투 등 200여종 (굴림, 신명조, 고딕체 사용)
- 명함, 기타 문서들 (신명조, 고딕, 궁서체 사용)

여러 관련 연구들에서 서체인식이 문서인식 및 정보검색의 자동화에 중요한 역할을 수행한다고 주장해 왔듯이, 실제 이러한 자료조사를 통해 서체인식의 필요성을 확인하였으며, 본 조사를 토대로 제안하는 서체 분류 시스템의 인식 대상은 명조, 고딕, 바탕, 돋움, 궁서, 굴림 서체의 6개 클래스로 결정하였다.

3. 서체 인식 시스템

본 연구에서는 한글 문자의 다양한 서체를 인식하기 위해서 다차원의 특징 및 복잡한 인식기를 구축하기보다는, 한글 문자를 크게 세리프와 산세리프 계열로 분류하는 문제와 그리고 분류된 각 클래스 안에서 보다 세밀한 서체 분류를 수행하는 계층적 접근 방법을 생각하고 있다. 이러한 계층적인 접근 방법에 의한 분할 해결법(divide-and-conquer)은 보다 적은 차원의 특징과 간소화된 인식기를 채택할 수 있다는 장점이 있다. 전체적인 시스템 개요는 그림 1에 잘 나타나 있다.

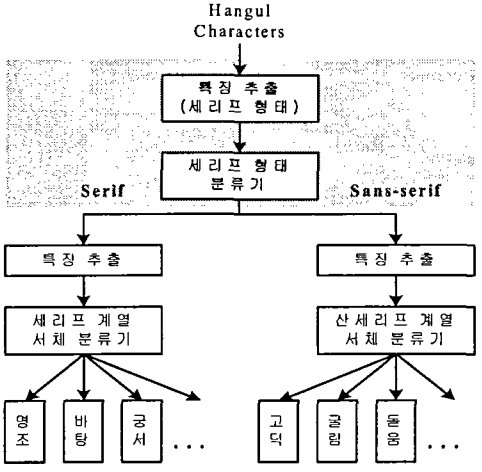

그림 1. 한글 서체 분류 시스템의 개념도

4. 세리프 영역의 방향 벡터

한글 문자의 서체 분류는 한글의 수직 획의 시작 부분에 나타나는 세리프의 형태에 따른 분

류이며, 세리프 영역에서 추출한 특징이 한글 문자의 서체 분류에 가장 효율적이라고 가정할 수 있다. 특징 추출은 먼저, 한글 문자의 세리프 영역을 추출하고 세리프 영역에 존재하는 런(run)들의 방향벡터를 추출하여 36등분면 상의 위치를 계산하였다.

4.1 세리프 영역 추출

한글의 세리프 영역은 수직 획의 시작 부분에 존재하며, 한글 영상에 대해 수평 및 수직 런 분석(run analysis)에 의한 세선화(thinning)를 적용해보면, 수평 런 분석에 의해 수직 획의 세그먼트(segment)를 구할 수 있고, 수직 런 분석을 통해 수평 획의 세그먼트를 구할 수 있으며, 수직 획과 수평 획이 만나는 연결영역을 구할 수 있다. 세리프 영역은 수직 획 세그먼트의 시작부분이 어떤 수평 획 세그먼트와도 만나지 않는 경우에 존재한다.

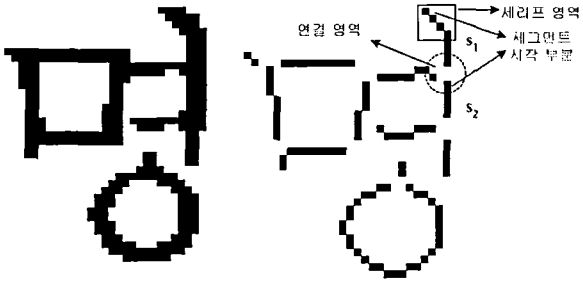

그림 2. 한글의 세리프 영역 추출

그림 2의 왼쪽은 입력된 한글 영상이고 오른쪽은 세선화된 영상을 나타낸다. 그림 2에서 세그먼트 s_1, s_2를 살펴보면, s_2는 시작 부분이 연결영역에 접해 있는 반면에 s_1은 그렇지 않다. 그러므로 세그먼트 s_1에서 5개의 런을 연결한 세리프 영역을 추출한다.

4.2 세리프 영역의 방향벡터

세리프 영역의 방향벡터는 영역에 존재하는 수평 런들의 방향벡터를 구한 후, 가리키는 방향을 36등분면 상의 위치로 결정한 값이다. 여기에서 수평 런들의 방향벡터는 이웃하는 런들의 중간점을 연결한 방향벡터로 나타낸다. 추출된 세리프 영역은 수평 런들의 중간점을 포함하고 있다. 그림 3에서 추출된 세리프 영역에 존재하는 5개 런들의 중간점을 보여주고 있다.

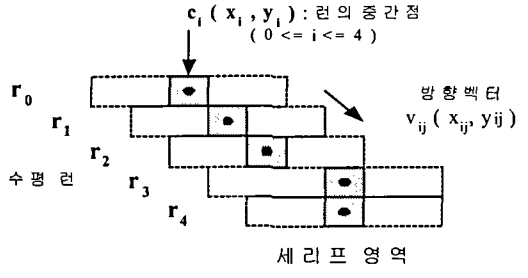

그림 3. 추출된 세리프 영역의 런

이웃하는 임의의 두 런들의 방향벡터 v_{ij}는 i번째 런에서 j번째 런으로 향하는 방향 벡터로, 다음과 같이 구할 수 있다.

$$v_{ij} = (x_j - x_i, y_j - y_i), \quad (j > i)$$

그리고 D_{ij}를 방향벡터의 36등분면 상에서 위치라고 하면, 세리프 영역의 방향벡터 D는 다음과 같이 모든 방향벡터의 평균을 취한다.

$$D = \left(\sum_{i=0}^{3} D_{ij} \right) / (N-1), \quad (j = i+1) \ (N : \text{세리프 영역에 있는 런의 개수})$$

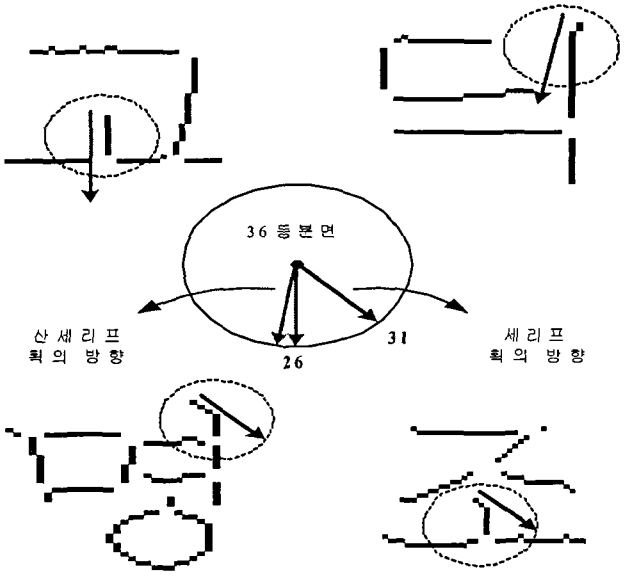

그림 4. 36등분면 상에서 방향벡터의 위치

그림 4는 한글의 세리프 영역에서 추출한 방향벡터와 36등분면 상의 위치를 나타내고 있다. 산세리프 계열의 한글은 획의 방향이 항상 수직으로 동일하지만, 세리프 계열은 획의 시작부분에 돌기가 있어서 돌기의 방향과 획의 방향이 상이함을 관찰할 수 있다.

5. 실험 결과 및 분석

본 논문에서는 한글 문자의 서체 분류를 위한 특징을 제안하였다. 제안한 특징의 유용성 검증을 위한 실험은 세리프 영역에서 추출한 방향벡터의 분포를 조사하였고, 서체인식기를 훈련하여 인식률을 측정하였다. 실험 데이터는 한글 문서 영상으로부터 분할한 낱자 3,000개이며, 그 구성은 다음과 같다.

- 세리프 계열 낱자 영상 : 1,500개 (명조, 바탕, 궁서, 각각 500개)
- 산세리프 계열 낱자 영상 : 1,500개 (고딕, 돋움, 굴림, 각각 500개)

그림 5에서 나타내고 있듯이, 산세리프 계열의 한글은 세리프가 존재하지 않기 때문에 세리프 영역의 방향이 전체 획의 방향과 일치하는 25~26사이에 분포가 주로 형성되었다. 반면, 세리프 계열의 한글은 세리프 영역의 방향과 획의 방향이 상이한 29~31 사이에 분포가 주로 형성되었다. 물론, 문서영상의 품질이 좋지 않아 한글의 세리프 영역이 심하게 훼손된 경우에는 원하는 결과를 얻지 못하는 경우도 있었다.

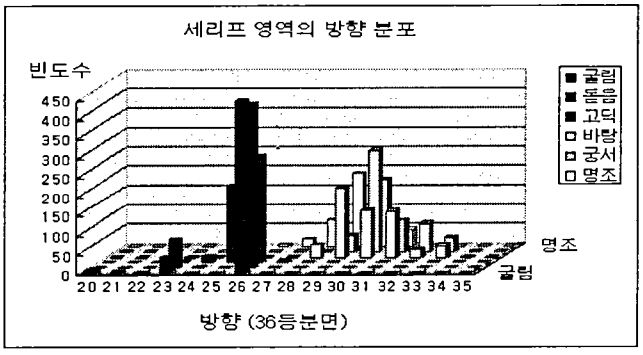

그림 5. 실험 결과의 히스토그램

실험을 통해 제안한 특징의 분별력을 확인하였는데, 세리프 계열의 경우에 영상의 품질에 따라 분포가 훨씬 넓게 나타남(그림 5)을 알 수 있었고, 인식 결과에서도 세리프 계열에 대한 성능이 더 낮게 나타났다. 또한, 한글에 수직 획이 전혀 존재하지 않는 경우('으', '그' 등)에 대해서는 처리 불가능하여 수평 획에의 적용이 불가피하였다.

6. 결론 및 향후 연구

본 논문에서는 한글 문자의 서체 분류 시스템을 소개하였고, 세리프와 산세리프 계열의 분류를 위한 특징을 제안하였다. 제안한 특징은 수직 획의 시작 부분에 위치하는 세리프 영역을 추출하고, 그 영역에 존재한 런들의 방향성을 36등분면 상의 위치 값으로 계산하였다. 그리고 문서영상으로부터 분할한 낱자 영상에 대한 실험을 통해 세리프 영역에서 추출한 방향 특징의 유용성을 입증하였다. 또한, 흔히 접할 수 있는 600여종의 문서들을 대상으로 자주 사용되는 서체에 대한 자료 조사를 통해 서체 인식의 필요성을 확인하였다.

향후 연구과제는 먼저, 수직 획이 존재하지 않는 경우에 대해 수평 획을 이용하는 방법 등에 대한 연구이다. 또한, 세리프(명조, 궁서, 바탕 등)와 산세리프(고딕, 굴림, 돋움 등) 클래스 안의 다양한 서체를 인식하기 위해 특징을 개발하고 전체적인 서체 분류 시스템을 구축하는 것이다.

참고문헌

[1] AIIM'96 Conference Handbooks, Association for Imaging and Information Methodologies, 1996.

[2] D. Doermann, "The Indexing and Retrieval of Document Images: A Survey," *Computer Vision and Image Understanding*, Vol. 70, No. 3, pp. 287-298, 1998.

[3] U. Garain and B.B. Chaudhuri, "Extraction of Type Style Based Meta-Information from Imaged Documents," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 341-344, 1999.

[4] H.S. Baird, G. Nagy, "A Self-Correcting 100 Font Classifier," *Proc. of SPIE Conference on Document Recognition*, pp. 106-115, 1994.

[5] A. Zramdini, *"Study of optical font recognition based on global typographical features,"* PhD thesis, University of Fribourg, 1995.

[6] S. Kahan, T. Pavlidis and H.S. Baird, "On the Recognition of Printed Characters of Any Font and Size," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 9, No. 2, pp. 274-288, 1987.

[7] M.C. Jung, Y.C. Shin and S.N. Srihari, "Multifont Classification Using Typographical Attributes," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 353-356, 1999.

[8] T.K. Ho, J.J. Hull and S.N. Srihari, "A Computational Model for Recognition of Multi-Font Images," *Machine Vision and Applications*, Vol. 5, No. 1, pp. 157-168, 1992.

[9] S. Zhao and S.N. Srihari, "A Word Recognition Algorithm for Machine-Printed Word Images of Multiple Fonts and Varying Qualities," *Proc. 3rd Int. Conf. Document Analysis and Recognition*, Montreal, pp. 351-354, 1995.

[10] B.B. Chaudhuri and U. Garain, "Automatic Detection of Italic, Bold and All-Capital Words in Document Images," *Proc. 14th Int. Conf. Pattern Recognition*, Brisbane, pp. 610-612, 1998.

[11] T.K. Ho, "Font Identification of Stop Words for Font Learning and Keyword Spotting", *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 333-336, 1999.

[12] Y. Zhu, T. Tan and Y. Wang, "Font Recognition Based on Global Texture Analysis," *Proc. 5th Int. Conf. Document Analysis and Recognition*, Bangalore, pp. 349-352, 1999.

[13] H. Shi, T. Pavlidis, "Font Recognition and Contextual Processing for More Accurate Text Recognition," *Proc. 4th Int. Conf. Document Analysis and Recognition*, Ulm, pp. 39-44, 1997.

[14] 박문호, 손영우, 김석태, 남궁재찬, "인쇄된 한글 문서의 폰트 인식," *한국정보처리논문지*, 제 4권, 제 8호, pp. 2017-2024, 1997.

[15] 곽희규, "*문서 영상의 단어 단위 분할 및 단어 영상의 속성 추출에 관한 연구*," 전남대학교 박사학위논문, 2001.

[16] J. Hochberg, L. Kerns, P. Kelly, and T. Thomas, "Automatic Script Identification from Images Using Cluster-based Templates," *Proc. 3rd Int. Conf. Document Analysis and Recognition*, Montreal, pp. 378-381, 1995.